Figure 8.11. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating an Average or a Total

1. Background Information

The mathematics on pages 8.52 and 8.53 in Section 5 of this Figure 8.11 is a centrepiece of the theory of survey sampling ('design-based' inference – recall Section 7 in Figure 6.1); we can appreciate it more fully with the following background information.

∗ In introductory probability, we use *upper* case *italic* letters (usually near the end of the alphabet, like X, Y and Z) for random variables and the corresponding *lower* cases letters (*e.g.*, x, y and z) for their values. We now further distinguish *upper*-case **bold** letters for *population* quantities; for example, $\mathbf{Y}_{\mathrm{i}}$ is the response variate for the i*th* element in the respondent population.

– The line through population symbols is make distinguishable *italic* and **bold** upper-case *hand*written letters and we *say*, for instance, '$\mathbf{Y}_{\mathrm{i}}$ cross' for the response variate of the i*th* respondent population element.

Table 8.11.1:

SYMBOL: Random variable	Value	Respondent population	DESCRIPTION
Y	y	$\mathbf{Y}$	Response variate
–	j	i	Summation index
–	x	$\mathbf{X}$	Focal explanatory variate
–	z	$\mathbf{Z}$	Explanatory variate
–	n	$\mathbf{N}$	Number of units/elements
$\overline{Y}$	$\overline{y}$	$\overline{\mathbf{Y}}$	Average (sum ÷ number)
R	r	$\mathbf{R}$	Residual [or Ratio]
S	s	$\mathbf{S}$	Standard deviation

∗ In introductory statistics courses, the number of elements in the population (also called the population *size*) is seldom considered explicitly; in this Figure 8.11, this attribute is denoted $\mathbf{N}$ ('**N** cross').

– Including the population size in survey sampling theory is sometimes called dealing with a *finite* population, but this is *unhelpful* terminology (perhaps carrying over from mathematics where infinity often arises). A population in statistics is a real-world entity and so, by its nature, has a finite number of elements – see also Note 2 on page 8.53.

∗ When investigating a Question with a *descriptive* aspect [one whose Answer will involve primarily values for population/process attributes (past, present, future)], a useful way to think of (or to 'model') the response variate of a respondent population element is as shown in equation (8.11.1) at the right;

$$\mathbf{Y}_{\mathrm{i}} = \overline{\mathbf{Y}} + \mathbf{R}_{\mathrm{i}} \qquad \text{-----(8.11.1)}$$

this model is useful because it expresses a quantity we can observe ($\mathbf{Y}_{\mathrm{i}}$) in terms of an attribute of interest ($\overline{\mathbf{Y}}$, the population *average*) and a quantity ($\mathbf{R}_{\mathrm{i}}$, the *population residual* for element i) whose behaviour is amenable to probability modelling which, ultimately, enables us to quantify imprecision due to sample error and (in some contexts) measurement error.

– The **response model** for a *non*-comparative Plan – the type of Plan usually appropriate to answer a Question with a descriptive aspect – involving equiprobable (simple random) selecting (EPS) of n units from an *un*stratified population is:

$$Y_j = \mu + R_j, \quad j = 1, 2, ..., \mathrm{n}, \quad R_j \sim N(0, \sigma), \quad \text{independent,} \quad \text{EPS,} \qquad \text{-----(8.11.2)}$$

where Y_j is a random variable whose distribution represents the possible values of the measured response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over;

R_j is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the measured value of the response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over.

The symbol μ in the model (8.11.2), and two related entities $\hat{\mu}$ and $\tilde{\mu}$, have the following meanings:

μ: a parameter representing the *average* ($\overline{\mathbf{Y}}$) of the measured response variate of the elements of the respondent population.

$\hat{\mu}$: the least squares *estimate* of μ – a *number* whose value is derived from an appropriate set of *data*;

- ○ for the model (8.11.2), $\hat{\mu} = \overline{y}$, reflecting the intuitive idea that, under EPS, the measured sample average estimates the measured respondent population average;

$\tilde{\mu}$: the least squares *estimator* of μ – a *random variable* whose distribution represents the possible values of the estimate $\hat{\mu}$ if the selecting, measuring and estimating processes were to be repeated over and over;

- ○ here, $\tilde{\mu} = \overline{Y}$, the random variable representing the sample average under EPS.

∗ We met σ (on page 5.6 in Figure 5.1), and two related entities $\hat{\sigma}$ and $\tilde{\sigma}$, in the context of response models like (8.11.2) above;

σ: the (probabilistic) *standard deviation* of the normal model for the distribution of the residual, is a model parameter representing the (data) *standard deviation* ($\mathbf{S}$) of the measured response variate of the respondent population elements; this (data) standard deviation (and, hence, σ) *quantifies* the *variation* of the measured response variate over the elements of the respondent population – as this variation increases, so does $\mathbf{S}$ (and, hence, so does σ).

- ○ two *other* characteristics of variation are its *location* and its *shape* – these are, respectively, 0 and normal for (8.11.2);

$\hat{\sigma}$: the least squares *estimate* of σ – a *number* whose value is derived from an appropriate set of *data*;

$\tilde{\sigma}$: the least squares *estimator* of σ – a *random variable* whose distribution represents the possible values of the estimate $\hat{\sigma}$ if the selecting, measuring and estimating processes were to be repeated over and over.

How σ is *estimated*, reflected by differing expressions for $\hat{\sigma}$, depends on the sampling protocol in the Plan for the investigation and the response model appropriate for this Plan; for the model (8.11.2), $\hat{\sigma}$ is given by equation (8.11.3).

$$\hat{\sigma} = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})^2} \qquad \text{-----(8.11.3)}$$

∗ In *this* Figure 8.11, we meet overleaf (on page 8.50) *four* quantities which are called '*standard deviation*'; the first two and an

(*continued overleaf*)

associated quantity S correspond to the three σs given overleaf on page 8.49 but, *un*like $\hat{\sigma}$, s is called a standard deviation.

- $\mathbf{S}$: the respondent *population* (data) standard deviation – it is defined in Section 4 in Table 8.11.5 on page 8.52 and is a *number* which quantifies the variation over the respondent population of the response variate $\mathbf{Y}$ about its average $\overline{\mathbf{Y}}$;
  - ○ like most population attributes except $\mathbf{N}$, usually the value of $\mathbf{S}$ is *un*known;
- s: the *sample* (data) standard deviation – it is defined in Table 8.11.5 on page 8.52 and is a *number* which quantifies the variation over the sample of the response y about its average $\overline{y}$; the expression for s is (8.11.3) at the bottom right overleaf on page 8.49, the *same* as that for $\hat{\sigma}$ in the model (8.11.2).
  - ○ under EPS, s is used to estimate $\mathbf{S}$ – that is, to provide a value we can use for $\mathbf{S}$;
  - ○ this Figure 8.11 is concerned with only *one* sampling protocol – EPS from an unstratified population to estimate an average or total – and so there is only *one* expression for the estimate (s) of $\mathbf{S}$;
- S: the *estimator* corresponding to s – under EPS, it is a *random variable*, of which s is one (realized) *value*;
- \+ $s.d.(\overline{Y})$: the standard deviation of the sample average – under EPS, it provides a theoretical basis for quantifying uncertainty due to sample error in estimates of respondent population attributes like an average or total;
- \+ $\widehat{s.d.}(\overline{Y})$: the *estimated* standard deviation of the sample average which, under EPS, is the basis for calculating *values* for the end points of confidence intervals for respondent population attributes like an average or total.

Table 8.11.2: SUMMARY OF STANDARD DEVIATIONS

Response Models	**Survey Sampling**
σ Model parameter	$\mathbf{S}$ Respondent population standard deviation – an attribute
$\hat{\sigma}$ Estimate of σ	s Sample standard deviation – an estimate of $\mathbf{S}$
$\tilde{\sigma}$ Estimator of σ	S Estimator corresponding to s – a random variable

The expressions for $s.d.(\overline{Y})$ and $\widehat{s.d.}(\overline{Y})$ differ in that $\mathbf{S}$ is replaced by its estimate s – see equation (8.11.9) and equation (8.11.17) on page 8.53. [In a non-probability sampling context concerned only with *data*, s would usually be denoted s.]

∗ We capitalize on having *two* words – average and mean – in English to make a useful distinction for a measure of *location*:
- the *average* is a measure of location for a set of *data*;
- the *mean* is a measure of location for (the distribution of) a *random variable.*

However, for the magnitude of *variation* there is only *one* term – standard deviation – for the commonly-used measure, and this can be a source of confusion. Ideally, we would like:
- the *(new word)* as a measure of variation for a set of *data*,
- the *standard deviation* as a measure of variation for a *random variable*,

but the use of 'standard deviation', regardless of context, is too well-established in statistics for this ideal to be attainable. A compromise, to assist beginning students, is to distinguish a *data* standard deviation from a *probabilistic* standard deviation – see Table 8.11.3 at the right. Note that *we* use one symbol (*e.g.*, $\mathbf{S}$, s) for a data standard deviation and the abbreviation *s.d.* for a probabilistic standard deviation.

Table 8.11.3

Respondent population standard deviation	$\mathbf{S}$	*data* standard deviation
Sample standard deviation	s	
Standard deviation of the sample average	$s.d.(\overline{Y})$	*probabilistic* standard deviation
Estimated standard deviation of the sample average	$\widehat{s.d.}(\overline{Y})$	

Figure 8.12 of these Materials helps us appreciate the distinction between the sample standard deviation (s; represented visually by the 16 'hooked' horizontal lines in each diagram) and the standard deviation of the sample average [$s.d.(\overline{Y})$; as estimated from the 16 sample averages and denoted $s_{\overline{y}}$ near the lower right-hand corner of each diagram].

∗ The standard deviation of $\overline{Y}$ is sometimes referred to as the *standard error* of $\overline{Y}$ (*e.g.*, Barnett, pp. 26, 45) but this term has been avoided in these Course Materials because it is used by different authors for *both* $s.d.(\overline{Y})$ and $\widehat{s.d.}(\overline{Y})$ (see also Cochran, pages 24, 25-27 and 53), potentially confusing a quantity and its estimate. [References are given on page 8.56 in Section 7.]

∗ The following suggestions may help avoid confusion arising from (careless use of) the terminology discussed above.
- ● when you encounter the word *mean*, be sure you understand whether it refers to:
  - an average of data (and whether the data are from a *sample* or a *census*), **OR**
  - a random variable [and whether it is an individual random variable or a (linear) combination (*e.g.*, an average, sum or difference)], **OR**
  - a parameter of a response model or probability model.
- ● when you encounter the term *standard deviation* or *standard error*, be sure you understand whether it refers to:
  - the variation of data (and whether the data are from a *sample* or a *census*), **OR**
  - a random variable [and whether it is an individual random variable or a (linear) combination (*e.g.*, an average, sum or difference)], **OR**
  - a parameter of a response model or probability model.
- ● when you encounter the word *inaccuracy*, remember that it is a real-world quantity and is defined only in the context of *repetition* of a *process* – like selecting or measuring.
  - *Estimating* bias (a model quantity) *differs* from inaccuracy in that it *de*creases with *in*creasing sample size and so may not be of much practical concern in actual sample surveys.

  [There is further discussion of bias in Appendix 3 and Appendix 4 on pages 8.57 and 8.58.]

(*continued*)

Figure 8.11. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating an Average or a Total (continued 1)

2. Equiprobable (Simple random) Selecting

The *definition* of equiprobable selecting (EPS) is: If a sample of n units is obtained from a respondent population of $\mathbf{N}$ units in such a way that every *sample* of size n has an *equal* probability of being selected, the selecting process is called **equiprobable selecting**. [Elsewhere, you may see it called **simple random selecting** (SRS).]

In practice, we often think of EPS as being *implemented* by selecting each *unit* of the sample equiprobably ('at random') and with*out* replacement ('EPSWOR') from the (unstratified) respondent population. [The element-unit distinction is discussed in Appendix 1 at the bottom of page 8.56 and the top of page 8.57.]

Because we commonly think of EPS in terms of how we select the *units*, we may overlook the fact that the definition is in terms of *sample* probabilities. In particular, we need to recognize that, while the definition implies that each *unit* has the same inclusion probability of $n/\mathbf{N}$, there are selecting processes with equal unit inclusion probabilities that are *not* EPS. An illustration is given at the right below; for this respondent population of $\mathbf{N} = 4$ units, six samples of size $n = 2$ can be obtained by EPS but only *two* such samples are obtained by *systematic* selecting; however, provided the starting point of the systematic selecting process is chosen equiprobably, any unit has an inclusion probability of ½ under *either* process.

Another way of making the same point is to say that, under systematic selecting, two of the six possible samples of size 2 have probability ½ and four have *zero* probability.

$\mathbf{N} = 4$ **Population units**
1, 2, 3, 4;
the samples of size 2 are:
EPS: (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4);
systematic selecting: (1, 3), (2, 4).

The emphasis in statistics on EPS (or its equivalent) is because it is the basis of theory which provides:

- *unbiased* estimating of a population average (an attribute commonly of interest);
- a connection between sampling imprecision and *sample size* (or level of *replicating*);
- an expression for a *confidence interval* for a population average – such an interval, under suitable modelling assumptions, *quantifies* sampling and measuring imprecision (as demonstrated in Figure 6.1 of these STAT 220 Course Materials).

[These three provisions of statistical theory refer to behaviour under *repetition* – Answer(s) obtained in a *particular* investigation *remain* uncertain, as reflected by their limitations.]

EPS does *not*, of itself, *reduce* sample error or sampling imprecision, as implied in (wrong) statements such as:

○ EPS generates a *representative* sample;
○ EPS generates a sample which provides a proper basis for *generalization*; !

as well as misrepresenting the statistical benefits of using EPS, such statements confuse repetition (the *process* of EPS) with a component of a *particular* investigation (the actual sample). A *correct* statement is:

EPS, *in conjunction with adequate replicating* (or an *adequate sample size*), provides for quantifying sampling imprecision and so allows a particular investigation to obtain an Answer with acceptable limitation due to sample error.

- What constitutes *acceptable* limitation depends on the investigation requirements for its Answer(s); for instance, in a poll to estimate one or more proportions, an acceptable limitation may be quantified as the proportion(s) estimated to within 2 percentage points 19 times out of 20. [Limitations may also be imposed by the *resources* available for the investigating].

3. Sample Size and Sample Error under EPS

Example 8.11.1: A respondent population of $\mathbf{N} = 4$ units has the following integer $\mathbf{Y}$-values for its response variate:

1, 2, 4, 5 (so that: $\overline{\mathbf{Y}} = 3$, $\mathbf{S} \simeq 1.8257$);

we examine the behaviour of *sample error* under EPS as the sample size *in*creases from 1 to 2 to 3 to 4.

The number at the bottom of the four error columns of Tables 8.11.4 below is the *average magnitude* of the sample error for that sample size.

Table 8.11.4a
EPS of n = 1 unit

Sample	$\bar{y}$	Error
(1)	1	−2
(2)	2	−1
(4)	4	1
(5)	5	2
		1½

Table 8.11.4b
EPS of n = 2 units

Sample	$\bar{y}$	Error
(1, 2)	1½	−1½
(1, 4)	2½	−½
(1, 5)	3	0
(2, 4)	3	0
(2, 5)	3½	½
(4, 5)	4½	1½
		⅔

Table 8.11.4c
EPS of n = 3 units

Sample	$\bar{y}$	Error
(1, 2, 4)	2⅓	−⅔
(1, 2, 5)	2⅔	−⅓
(1, 4, 5)	3⅓	⅓
(2, 4, 5)	3⅔	⅔
		½

Table 8.11.4d
EPS of n = 4 units

Sample	$\bar{y}$	Error
(1, 2, 4, 5)	3	0
		0

Example 8.11.1 reminds us of general results under EPS that follow from the theory in Section 5 on pages 8.52 and 8.53.

- as the sample size *in*creases, the average magnitude (and, hence, the standard deviation) of sample error *de*creases – this is what we mean when we say that increasing sample size *de*creases sampling *imprecision* under EPS;

(*continued overleaf*)

- taking the *sign* of sample error into account, the average error is *zero* for each n – this is what we mean by saying that, under EPS, (the random variable representing) the sample average is an *unbiased* estimator of the respondent population average;
  - note that *both* the selecting method *and* the population attribute and its estimator are involved in this statement;
  - another statement with these components, which contrasts with the statement above about $\overline{Y}$, is that, for the population attribute which is the *ratio* of the average of two response variates ($\mathbf{R} = \overline{\mathbf{Y}}/\overline{\mathbf{X}}$), the sample ratio $r = \overline{y}/\overline{x}$ is *biased* $[E(R) \neq \mathbf{R}]$ under EPS but *un*biased if the first sample unit is selected with probability proportional to its $\mathbf{X}$ value and the remainder selected equiprobably (see Cochran, p. 175).
- there is no *sample* error when a *census* is taken – when *all* units of the respondent population are selected.

We also see that there can be a sample size(s) for which *none* of its $\binom{\mathbf{N}}{n}$ samples has *zero* sample error – *no* sample has $\overline{y} = \overline{\mathbf{Y}}$.

4. Notation

Table 8.11.5 below gives the notation used in the theory developed in this Figure 8.11; the last column of the table includes the 'model'. It is a model only in the sense of being an *idealization* or *mathematical abstraction* involving the equal probabilities attained under EPS; it is *not* a model in the sense of a symbolic expression like a response model [such as equation (8.11.2) on the first side (page 8.49) of this Figure 8.11].

Table 8.11.5: ....QUANTITY......	**RESPONDENT POPULATION**	.......**SAMPLE [MODEL]**.....................
Size (elements/units)	$\mathbf{N}$	n
Response	$\mathbf{Y}_i$ (i = 1, 2, ..., $\mathbf{N}$)	y_j (j = 1, 2, ..., n) [r.v.s are Y_j]
Average	$\overline{\mathbf{Y}} = \frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i = \frac{1}{\mathbf{N}}\,{}_{T}\mathbf{Y}$	$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ [r.v. is $\overline{Y}$]
Total	${}_{T}\mathbf{Y} = \mathbf{N}\overline{\mathbf{Y}} = \sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i$	${}_{T}y = \mathbf{N}\overline{y}$ [r.v. is ${}_{T}Y$]
Standard deviation	$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}SS_{\mathbf{Y}}} \equiv \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^2}$	$s = \sqrt{\frac{1}{n-1}SS_y} \equiv \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2}$ [r.v. is S]

${}_{T}y$, the *estimate* of the population total ${}_{T}\mathbf{Y}$, is *not* the *sample* total, which is $n\overline{y} = \sum_{j=1}^{n} y_j$ and is usually *not* a sample attribute of interest.

5. Estimating $\overline{\mathbf{Y}}$, the Respondent Population Average

We want both a value (or *point estimate*) for this respondent population attribute *and* a measure of the (sampling) uncertainty of the estimate, for which we use a confidence interval.

To develop the relevant theory, we first establish results for $E(Y_j)$, $s.d.(Y_j)$ and $cov(Y_j, Y_l)$:

(i) $E(Y_j) = \sum_{i=1}^{\mathbf{N}} \mathbf{Y}_i \cdot \Pr(Y_j = \mathbf{Y}_i)$ (the mean of a discrete random variable).

We can find $\Pr(Y_j = \mathbf{Y}_i)$ in any of three ways:

(a) because every possible ordered sample is equally probable under equiprobable selecting, any population unit is equally probable at any position in the sample and, because there are $\mathbf{N}$ units in the population, this probability is $1/\mathbf{N}$;

(b) ordered counting: $\dfrac{\text{number of ordered samples with } Y_j = \mathbf{Y}_i}{\text{total number of samples of size n}} = \dfrac{(\mathbf{N}-1)^{(n-1)}}{\mathbf{N}^{(n)}} = \dfrac{1}{\mathbf{N}};$ -----(8.11.4)

(c) *un*ordered counting: $\dfrac{\text{number of } un\text{ordered samples with } \mathbf{Y}_i \text{ at any position}}{\text{total number of samples of sze n}} \cdot \dfrac{1}{\text{number of sample positions}} = \left[\binom{\mathbf{N}-1}{n-1}\Big/\binom{\mathbf{N}}{n}\right] \cdot [1/n] = \dfrac{1}{\mathbf{N}};$

$\therefore \quad E(Y_j) = \sum_{i=1}^{\mathbf{N}} \mathbf{Y}_i \cdot \frac{1}{\mathbf{N}} = \frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}} \mathbf{Y}_i = \overline{\mathbf{Y}}.$ -----(8.11.5)

(ii) $E(Y_j^2) = \sum_{i=1}^{\mathbf{N}} \mathbf{Y}_i^2 \cdot \Pr(Y_j = \mathbf{Y}_i) = \frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}} \mathbf{Y}_i^2$ [using the result for $\Pr(Y_j = \mathbf{Y}_i)$ from (**i**)];

$\therefore \quad s.d.(Y_j) = \sqrt{E(Y_j^2) - [E(Y_j)]^2} = \sqrt{\frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 - \overline{\mathbf{Y}}^2} = \sqrt{\frac{1}{\mathbf{N}}\Big[\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 - \mathbf{N}\overline{\mathbf{Y}}^2\Big]} = \sqrt{\frac{\mathbf{N}-1}{\mathbf{N}}}\,\mathbf{S}.$ -----(8.11.6)

(iii) $E(Y_j Y_l) = \sum_{i=1}^{\mathbf{N}} \sum_{k \neq i=1}^{\mathbf{N}} \mathbf{Y}_i \mathbf{Y}_k \cdot \Pr(Y_j = \mathbf{Y}_i, Y_l = \mathbf{Y}_k)$ (the mean of a product of discrete random variables).

But from (**i**), because $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B|A)$, $\Pr(Y_j = \mathbf{Y}_i, Y_l = \mathbf{Y}_k) = \Pr(Y_j = \mathbf{Y}_i) \cdot \Pr(Y_l = \mathbf{Y}_k | Y_j = \mathbf{Y}_i) = \frac{1}{\mathbf{N}} \cdot \frac{1}{\mathbf{N}-1};$

$\therefore \quad E(Y_j Y_l) = \frac{1}{\mathbf{N}} \cdot \frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i \sum_{k \neq i=1}^{\mathbf{N}}\mathbf{Y}_k = \frac{1}{\mathbf{N}} \cdot \frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i(\mathbf{N}\overline{\mathbf{Y}} - \mathbf{Y}_i) = \frac{1}{\mathbf{N}} \cdot \frac{1}{\mathbf{N}-1}\Big[\mathbf{N}\overline{\mathbf{Y}}^2 - \sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2\Big]$ (because $\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i = \mathbf{N}\overline{\mathbf{Y}}$);

$\therefore \quad cov(Y_j, Y_l) = E(Y_j Y_l) - E(Y_j) \cdot E(Y_l)$

$$= \frac{1}{\mathbf{N}(\mathbf{N}-1)}\Big[\mathbf{N}^2\overline{\mathbf{Y}}^2 - \sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2\Big] - \overline{\mathbf{Y}} \cdot \overline{\mathbf{Y}} = \frac{\mathbf{N}^2\overline{\mathbf{Y}}^2 - \sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 - \mathbf{N}(\mathbf{N}-1)\overline{\mathbf{Y}}^2}{\mathbf{N}(\mathbf{N}-1)} = \frac{-\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 + \mathbf{N}\overline{\mathbf{Y}}^2}{\mathbf{N}(\mathbf{N}-1)} = -\frac{\mathbf{S}^2}{\mathbf{N}}. \quad \text{-----(8.11.7)}$$

(*continued*)

Figure 8.11. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating an Average or a Total (continued 2)

Hence, when we use $\overline{Y}$ (the random variable representing the sample average) as an estimator of $\overline{\mathbf{Y}}$ (the respondent population average), under EPS we have:

$$E(\overline{Y}) = E[\tfrac{1}{n}\sum_{j=1}^{n} Y_j] = \tfrac{1}{n}E[\sum_{j=1}^{n} Y_j] = \tfrac{1}{n}\sum_{j=1}^{n} E(Y_j) = \tfrac{1}{n}[\underbrace{\overline{\mathbf{Y}} + \overline{\mathbf{Y}} + \ldots\ldots + \overline{\mathbf{Y}}}_{\text{n terms from (i)}}] = \overline{\mathbf{Y}};$$

i.e., $\overline{Y}$ is an *unbiased* estimator of $\overline{\mathbf{Y}}$ under EPS; -----(8.11.8)

$$s.d.(\overline{Y}) = s.d.[\tfrac{1}{n}\sum_{j=1}^{n} Y_j] = \tfrac{1}{n}s.d.[\sum_{j=1}^{n} Y_j]$$

$$= \tfrac{1}{n}\sqrt{\sum_{j=1}^{n}[s.d.(Y_j)]^2 + \sum_{j=1}^{n}\sum_{l\neq j=1}^{n} cov(Y_j, Y_l)} = \tfrac{1}{n}\sqrt{\underbrace{n(\mathbf{N}-1)\mathbf{S}^2/\mathbf{N}}_{\text{n equal terms from (ii)}} + \underbrace{n(n-1)[-\mathbf{S}^2/\mathbf{N}]}_{n^2-n\text{ equal terms from (iii)}}} = \mathbf{S}\sqrt{\tfrac{1}{n} - \tfrac{1}{\mathbf{N}}} \qquad\text{-----(8.11.9)}$$

[the standard deviation of the sample average under EPS (from an *un*stratified respondent population)].

Thus, the distribution of (the estimator of) the sample average under EPS is: $\overline{Y} \dot{\sim} N(\overline{\mathbf{Y}}, \mathbf{S}\sqrt{\tfrac{1}{n} - \tfrac{1}{\mathbf{N}}})$ -----(8.11.10)

NOTES: 1. Equation (8.11.9) for $s.d.(\overline{Y})$ shows that the effect of the *finite* size of the respondent population is to modify the familiar expression by including a second term, $1/\mathbf{N}$, under the square root multiplying $\mathbf{S}$. Other matters of interest are:

- when $n = \mathbf{N}$, $s.d.(\overline{Y}) = 0$, reminding use that sample error is *zero* in a census.
- when $n \ll \mathbf{N}$ (*i.e.*, when the sample size is a small proportion of the respondent population size, say 5% or less), the expression for $s.d.(\overline{Y})$ becomes essentially the more familiar form $\mathbf{S}\sqrt{\tfrac{1}{n}}$. -----(8.11.11)
- the form of the square root multiplying $\mathbf{S}$ means that the precision of estimating $\overline{\mathbf{Y}}$ by $\overline{Y}$ under EPS is determined primarily by the *sample* size and only weakly by the *population* size.
  - This insight of statistical theory is counter-intuitive – there is essentially the *same* sampling imprecision in a national poll of 1,500 people selected from a population of 30 million Canadians or 300 million Americans.

2. The expression (8.11.9) may be written as shown in equations (8.11.12) and (8.11.13) at the right. The former, where $\mathbf{S}$ has been replaced by its expression in terms of $\mathbf{Y}$, is of interest to compare with equation (8.11.14) below; equation (8.11.13) gives the standard deviation of $\overline{Y}$ as the familiar form (8.11.11) multiplied by the square root of a *finite population correction* $1-f$, where $f = n/\mathbf{N}$ is the *sampling fraction.* BUT:

$$s.d.(\overline{Y}) = \sqrt{\tfrac{1}{\mathbf{N}-1}(\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^2} \qquad\text{-----(8.11.12)}$$

$$s.d.(\overline{Y}) = \sqrt{(1-f)}\,\mathbf{S}\sqrt{\tfrac{1}{n}} \qquad\text{-----(8.11.13)}$$

- Thinking of $s.d.(\overline{Y})$ as an 'infinite population' result times a 'correction factor' unhelpfully encourages confusing a *model* with the real world – recall the comment below Table 8.11.1 at the end of the second asterisk (*) on page 8.49.

3. The *coefficient of variation* (*c.v.*) of $\overline{Y}$ [a measure of *relative* imprecision] is given in equation (8.11.14) at the right. Relative imprecision *de*creases [*i.e.*, $s.d.(\overline{Y})$ becomes *smaller relative to* $\overline{\mathbf{Y}}$] as n becomes larger and when the $\mathbf{Y}_i$ have less variation about their average $\overline{\mathbf{Y}}$.

$$c.v.(\overline{Y}) \equiv \frac{s.d.(\overline{Y})}{\overline{\mathbf{Y}}} = \sqrt{\tfrac{1}{\mathbf{N}-1}(\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})\sum_{i=1}^{\mathbf{N}}(\frac{\mathbf{Y}_i}{\overline{\mathbf{Y}}} - 1)^2} \qquad\text{-----(8.11.14)}$$

4. $\overline{Y}$ is the linear unbiased estimator of $\overline{\mathbf{Y}}$ with smallest standard deviation based on a sample of size n units selected by EPS (*e.g.*, see Barnett, pp. 26-27).

5. The expression (8.11.9) for $s.d.(\overline{Y})$ under EPS is useful in three ways:
   - it gives the imprecision of the estimator $\overline{Y}$;
   - it allows us to calculate the approximate sample size needed to attain a specified imprecision for estimating $\overline{\mathbf{Y}}$ – recall Section 5 on page 6.26 in Figure 6.3 of these STAT 220 Course Materials;
   - it allows us to compare the efficiency of $\overline{Y}$ with that of *other* estimators of $\overline{\mathbf{Y}}$.

A practical difficulty in using the expression (8.11.9) above for $s.d.(\overline{Y})$ is that $\mathbf{S}$ is usually *un*known; a way around this difficulty is to use the *sample* standard deviation, s, as an estimate of $\mathbf{S}$, ostensibly because of the following:

$$s^2 = \frac{1}{n-1}[\sum_{j=1}^{n} y_j^2 - n\overline{y}^2] = \frac{1}{n-1}\sum_{j=1}^{n} y_j^2 - \frac{n}{n-1}\overline{y}^2; \qquad\text{-----(8.11.15)}$$

$$\therefore\ E(S^2) = \frac{1}{n-1}E[\sum_{j=1}^{n} Y_j^2] - \frac{n}{n-1}E(\overline{Y}^2) = \frac{1}{n-1}\underbrace{n\frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2}_{\text{n equal terms from (ii)}} - \frac{n}{n-1}\{\overline{\mathbf{Y}}^2 + [s.d.(\overline{Y})]^2\}$$

$$= \frac{n}{n-1}\{\frac{1}{\mathbf{N}}(\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 - \mathbf{N}\overline{\mathbf{Y}}^2) - (\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})\mathbf{S}^2\}$$

$$= \frac{n}{n-1}\{\frac{\mathbf{N}-1}{\mathbf{N}}\mathbf{S}^2 - (\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})\mathbf{S}^2\} = \mathbf{S}^2; \qquad\text{-----(8.11.16)}$$

$\{[s.d.(\overline{Y})]^2 = E(\overline{Y}^2) - [E(\overline{Y})]^2$, and: $E(\overline{Y}) = \overline{\mathbf{Y}}$ so that $E(\overline{Y}^2) = \overline{\mathbf{Y}}^2 + [s.d.(\overline{Y})]^2\}$

i.e., S^2 [the random variable representing the *square* of the sample (data) standard deviation under equiprobable selecting] is an *un*biased estimator of $\mathbf{S}^2$, the *square* of the respondent population (data) standard deviation [but see Appendix 3 on page 8.57].

Thus, the *estimated* standard deviation of $\overline{Y}$ under EPS is given by: $\hat{s.d.}(\overline{Y}) = s\sqrt{\tfrac{1}{n} - \tfrac{1}{\mathbf{N}}}$ -----(8.11.17)

On the basis of the approximate normality of the distribution of $\overline{Y}$ as a consequence of the Central Limit Theorem, and arguing in a general way from the use of the t distribution in normal theory when the population standard deviation is estimated by the sample standard deviation, the theory developed above leads to a probabilistic interval:

$$I = [\overline{Y} - {}_{\alpha}t^*_{\mathrm{n}-1} S\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}},\ \overline{Y} + {}_{\alpha}t^*_{\mathrm{n}-1} S\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}] \qquad \text{-----(8.11.18)}$$

such that $\Pr(I \ni \overline{\mathbf{Y}}) \simeq 100(1-\alpha)\%$, where ${}_{\alpha}t^*_{\mathrm{n}-1}$ is the $100(1-\alpha/2)th$ percentile of the $t_{\mathrm{n}-1}$ distribution. For calculating an approximate $100(1-\alpha)\%$ *confidence interval* for $\overline{\mathbf{Y}}$, we use:

$$\overline{y} \pm {}_{\alpha}t^*_{\mathrm{n}-1} s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = [\overline{y} - {}_{\alpha}t^*_{\mathrm{n}-1} s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}},\ \overline{y} + {}_{\alpha}t^*_{\mathrm{n}-1} s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}]. \qquad \text{-----(8.11.19)}$$

NOTES: 6. The *first* expression in (8.11.19) is convenient when assessing sampling imprecision; the *second* is a more *direct* Answer.

7. The values of sample attributes determine two characteristics of the approximate confidence interval for $\overline{\mathbf{Y}}$ – the sample average ($\overline{y}$) defines its *centre*, the sample standard deviation (s) determines its *width*; *both* characteristics (centre *and* width) of a confidence interval may be adversely affected by *inaccurate* selecting or measuring processes.

8. Using the t distribution requires that the population unit responses be *probabilistically independent* and *normally* distributed; in equiprobable selecting, successive observations are (weakly) dependent and not (necessarily) normally distributed, so this use of the t distribution has a weakened theoretical basis.
   - This weaker theoretical basis is one reason why the confidence interval expressions (8.11.19) are approximate.

9. An important consideration in assessing the nominal *level* of a confidence interval is how large the sample size needs to be for reasonable normality of $\overline{Y}$ as a consequence of the Central Limit Theorem; unfortunately, there is no reliable general rule but, when the deviation from normality is mainly a positive skewness, a crude rule (see Cochran, page 42) which is occasionally useful is that n should be greater than $25G_1^2$, where:

$$G_1 = \frac{1}{\mathbf{N}\mathbf{S}^3}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^3, \qquad \text{[estimated from the sample as: } g_1 = \frac{1}{\mathrm{n}s^3}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})^3\text{].} \qquad \text{-----(8.11.20)}$$

   - The approximate normality of the distribution of $\overline{Y}$ is a second (related) reason why the confidence interval expressions (8.11.19) are approximate.

10. The results derived in (**i**), (**ii**) and (**iii**) on the fourth side (page 8.52) of this Figure, which provide the theoretical basis for the confidence interval expressions, all involve the equal *unit* selection probabilities that are a consequence of EPS and, in (**iii**), the *joint* probability $1/\mathbf{N}(\mathbf{N}-1)$, which comes from the formal requirement for equiprobable *samples* under EPS. There is thus ***no*** basis for using these expressions to calculate a confidence interval from a sample obtained by *other* selecting methods (accessibility, haphazard, judgement, quota, systematic, volunteer, etc.).
   - Likewise, use throughout this Figure 8.11 of lower-case *italic y*s (values of random variables) to represent the measured sample response variate *data* values is based on EPS as the sample selecting process; other (*non*-probability) selecting processes would entail using instead Roman ys to represent such data values and there would be no reasonable basis for treating these ys as the *y*s of the foregoing theory (recall the comment in Figure 6.1 at the top of page 6.4 and Note 11 at the top of page 6.28 of Figure 6.3).

11. The theory leading to equation (8.11.9) overleaf on page 8.53 considers *only* sample error but using equation (8.11.17) when calculating a confidence interval for $\overline{\mathbf{Y}}$ or ${}_{\mathrm{T}}\mathbf{Y}$ involves using the *measured* sample y_js to calculate s. As a consequence, the confidence expressions (8.11.19) above for $\overline{\mathbf{Y}}$ and (8.11.23) near the bottom of the facing page 8.55 for ${}_{\mathrm{T}}\mathbf{Y}$ quantify *both* sample error *and* (fortuitously) measurement error.
   - We see that this is so by considering a respondent population whose elements all have the *same* $\mathbf{Y}$ value; variation in the y_js would then reflect *only* measurement error. Hence, in the usual case of *varying* $\mathbf{Y}_i$s, the measured y_j values reflect both sample and measurement error.
   - The confidence interval expressions (8.11.19) above and (8.11.23) on page 8.55 quantify the *combined* uncertainty due to sample error and measurement error – their effects could be estimated *individually* if replicate measurements were to be made on the sample units, but this is rare in sample surveys because there would be little benefit, extra cost and the difficulty of maintaining (real-world) *independence* of replicate measurements, especially when the population elements are humans and the measuring instrument is a questionnaire.

     This matter is the survey sampling analogue of the theory developed in Figure 6.1 for the model (8.11.2) on page 8.49 – for example, recall equation (6.1.18) on page 6.6.
   - It would be useful if the ('finite population') theory in *this* Figure 8.11 could inform that of Figure 6.1 so that it would be correct, when the population size is $\mathbf{N}$, to write equation (6.1.18) on page 6.6 as equation (8.11.21).

$$\overline{Y} \sim N(\mu, \sigma\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}) \qquad \text{----(8.11.21)}$$

Example 8.11.2: In an audit of hospital accounts, 200 accounts were obtained by equiprobable selecting from a total of 1,000 accounts; for all 200 accounts, the sample average was $\overline{y} = \$392.42$ and the sample standard deviation was $s = \$20.11$. Find an approximate 90% confidence interval for the *average* amount per account ($\overline{\mathbf{Y}}$) at the hospital.

(*continued*)

Figure 8.11. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating an Average or a Total (continued 3)

Solution: **NOTES:** 12. In calculating the 90% confidence interval for $\bar{\mathbf{Y}}$, the value of ${}_{.1}t^*_{199} = 1.65255$ has been obtained by *linear interpolation* from the relevant entries (*viz*., 1.65291 and 1.65251) from Table 6.4 (pages 6.33 and 6.34) for 190 and 200 degrees of freedom.

13. In calculations like those in the solution of Example 8.11.2, we must use and show enough significant figures to avoid rounding inaccuracy; however, it is an *essential* part of a proper solution to give a *final* answer rounded to a number of figures appropriate to the Question context.

Example 8.11.3: In the same hospital as in Example 8.11.2, n = 9 accounts were obtained by equiprobable selecting from the total of 484 open accounts; the data, and their numerical summaries, were as follows:

\$333.50 332.00 352.00 343.00 340.00 341.00
\$345.00 342.50 339.00

$$\sum_{j=1}^{9} y_j = 3{,}068.00, \qquad \sum_{j=1}^{9} y_j^2 = 1{,}046{,}132.50.$$

Find an approximate 95% confidence interval for the average amount per open account at the hospital.

Solution: The solution of this Example 8.11.3 is like that of Example 8.11.2 except *we* must calculate the values of $\bar{y}$ and s from the numerical summaries of the sample data.

We have: $\mathbf{N} = 484$, $\quad n = 9$, $\quad \bar{y} = \frac{3,068.00}{9} = \$340.\dot{8}$, $\quad {}_{.05}t^*_8 = 2.30600$ for 95% confidence,

$$s = \sqrt{\frac{1{,}046{,}132.50 - 3{,}068.00^2/9}{8}} = \$5.972\,739.$$

Then: $s\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} = 5.97274\sqrt{\frac{1}{9} - \frac{1}{484}} = \$1.972\,315\,622.$

Hence, an approximate 95% confidence interval for $\bar{\mathbf{Y}}$, the average amount per open account at the hospital, is:

$\bar{y} \pm 2.30600 \times \hat{sd}.(\bar{Y}) = 340.\dot{8} \pm 2.30600 \times 1.972\,316 \Rightarrow (336.34,\ 345.44)$ or about (\$336, \$346).

NOTES: 14. Despite the small sample size of 9, the confidence interval is, as in Example 8.11.2, relatively *narrow* (*i.e*., the Answer shows relatively *low* imprecision for estimating $\bar{\mathbf{Y}}$) because the value of the population standard deviation $\mathbf{S}$ is *small* in relation to the value of $\bar{\mathbf{Y}}$, as indicated by their estimates from the sample of about \$6 for s and about \$340 for $\bar{y}$.

- The small sample size in Example 8.11.3, where *all* the sample data are given, is only for classroom convenience; a real sample survey like this would usually have a *much* larger sample size.

15. For interest, we can carry out the check discussed in Note 9 on the facing page 8.54 for the adequacy of the sample size with respect to the assumed normality of $\bar{Y}$; for convenience, estimating the sum of cubes from the sample data is set out in Table 8.11.6 at the right; dividing the sum of cubes by $ns^3 = 1{,}917.622\,626$, we find $25g_1^2 = 4.453\,978$, which *is* less than n = 9 as the check requires.

Table 8.11.6:

j	y_j	$\bar{y}$	$y_j - \bar{y}$	$(y_j - \bar{y})^3$
1	333.50	$340.\dot{8}$	$-7.3\dot{8}$	−403.401 405
2	332.00	$340.\dot{8}$	$-8.8\dot{8}$	−702.331 959
3	352.00	$340.\dot{8}$	$11.\dot{1}$	1,371.742 116
4	343.00	$340.\dot{8}$	$2.\dot{1}$	9.408 779
5	340.00	$340.\dot{8}$	$-0.\dot{8}$	−0.702 332
6	341.00	$340.\dot{8}$	$0.\dot{1}$	0.001 372
7	345.00	$340.\dot{8}$	$4.\dot{1}$	69.482 854
8	342.50	$340.\dot{8}$	$1.6\dot{1}$	4.181 927
9	339.00	$340.\dot{8}$	$-1.\dot{8}$	−6.739 369
				341.641 983

6. Estimating ${}_{T}\mathbf{Y}$, the Respondent Population Total

Under the assumption that the population size, $\mathbf{N}$, is a *known constant*, the theory of equiprobable selecting for estimating ${}_{T}\mathbf{Y}$ is a straight-forward extension of the results for $\bar{\mathbf{Y}}$. Because the population total is ${}_{T}\mathbf{Y} = \mathbf{N}\bar{\mathbf{Y}}$, its estimator is $\mathbf{N}\bar{Y}$; the standard deviation of this estimator is then $\mathbf{N} \times s.d.(\bar{Y})$. Hence, we obtain a probabilistic interval:

$$I = [\mathbf{N}\bar{Y} - {}_{\alpha}t^*_{n-1}\mathbf{N}S\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}},\ \mathbf{N}\bar{Y} + {}_{\alpha}t^*_{n-1}\mathbf{N}S\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}}] \qquad \text{-----(8.11.22)}$$

such that $\Pr(I \ni {}_{T}\mathbf{Y}) \simeq 100(1-\alpha)\%$, where ${}_{\alpha}t^*_{n-1}$ is the $100(1-\alpha/2)th$ percentile of the t_{n-1} distribution. For calculating an approximate $100(1-\alpha)\%$ *confidence interval* for ${}_{T}\mathbf{Y}$, we use:

$$\mathbf{N}\bar{y} \pm {}_{\alpha}t^*_{n-1}\mathbf{N}s\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} = [\mathbf{N}\bar{y} - {}_{\alpha}t^*_{n-1}\mathbf{N}s\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}},\ \mathbf{N}\bar{y} + {}_{\alpha}t^*_{n-1}\mathbf{N}s\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}}]. \qquad \text{-----(8.11.23)}$$

Example 8.11.4: A company was concerned about the time per week its 750 managers spent on unimportant tasks. For 50 managers obtained by equiprobable selecting, it was found that the average time spent on such tasks was 10.31 hours and the standard deviation was 1.5 hours. Find an interval estimate for the total person-hours spent per week by the 750 managers on these unimportant tasks.

(*continued overleaf*)

Solution: Estimating ${}_{\mathrm{T}}\mathbf{Y}$ is like estimating $\overline{\mathbf{Y}}$ except that we must multiply by $\mathbf{N}$ in appropriate places; the solution of this Example 8.11.4 therefore follows the pattern of Examples 8.11.2 and 8.11.3.

We have: $\mathbf{N} = 750$, $\mathrm{n} = 50$, $\bar{y} = 10.31$ hours, $s = 1.5$ hours, ${}_{.05}t^*_{49} = 2.00958$ for 95% confidence.

Then: $s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 1.5\sqrt{\frac{1}{50} - \frac{1}{750}} = 0.204\,939$ hours.

Hence, an approximate 95% confidence interval for ${}_{\mathrm{T}}\mathbf{Y}$, the total number of hours spent by the 750 managers on unimportant tasks, is:

$$\mathbf{N}\bar{y} \pm 2.00958 \times \mathbf{N} \times \hat{s.d.}(\overline{Y}) = 750 \times 10.31 \pm 2.00958 \times 750 \times 0.204\,939 \implies (7{,}424,\ 8{,}041)$$

or about (7,400, 8,100) hours per week.

NOTES: 16. When the confidence *level* is not specifed in the Question, we take the default as 95%.

17. Population *totals* are often *large* numbers and so the widths of confidence intervals for ${}_{\mathrm{T}}\mathbf{Y}$ may be large in *absolute* terms but not necessarily large *relative* to the magnitude of ${}_{\mathrm{T}}\mathbf{Y}$.

18. It would be difficult to implement an accurate and precise measuring system for quantifying personal time usage for activities like those in Example 8.11.4; this is why the end points of the final confidence interval have been rounded to only *two* significant digits.

Example 8.11.5: One hundred water meters, obtained by equiprobable selecting from a community of 10,000 households, are monitored over a particular dry spell of weather. For all 100 meters, the sample average and standard deviation (in suitable units) are found to be 12.5 and 35.4 respectively. Find an approximate 99% confidence interval for the *total* water consumption in the community during the dry spell.

Solution: We have: $\mathbf{N} = 10{,}000$, $\mathrm{n} = 100$, $\bar{y} = 12.5$ units, $s = 35.4$ units, ${}_{.01}t^*_{99} = 2.62641$ for 99% confidence.

Then: $s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 35.4\sqrt{\frac{1}{100} - \frac{1}{10{,}000}} = 3.522\,256$ units.

Hence, an approximate 99% confidence interval for ${}_{\mathrm{T}}\mathbf{Y}$, the total water consumption of the 10,000 households over the dry spell, is:

$$\mathbf{N}\bar{y} \pm 2.62641 \times \mathbf{N} \times \hat{s.d.}(\overline{Y}) = 10{,}000 \times 12.5 \pm 2.62641 \times 10{,}000 \times 3.522\,256 \implies (32{,}491,\ 217{,}509)$$

or about (32,000, 220,000) units.

NOTE: 19. The *wide* confidence interval (*i.e.*, the high *imprecision*) for estimating ${}_{\mathrm{T}}\mathbf{Y}$ in Example 8.115 is mainly a consequence of a very *variable* population of household water consumptions: $s/\bar{y} = 283\%$. Because water consumption is an inherently *non*-negative quantity, these sample attribute values suggest a highly (positively) skewed population distribution of water consumptions which, in turn, raises concerns about the accuracy of the nominal confidence *level* of an interval based on the t distribution. In a real sample survey, this matter would need to be followed up.

- The high imprecision for estimating ${}_{\mathrm{T}}\mathbf{Y}$ in Example 8.11.5 could be managed by *stratifying* the population into groups of households more *homogeneous* with respect to their water consumptions; stratifying is discussed briefly in Appendix 5 on the last side (page 6.12) of Figure 6.1 and is pursued in more detail in Part 4 of the STAT 332 Course Materials.

7. REFERENCES:

1. Barnett, V. *Sample Survey Principles and Methods.* Second edition, Edward Arnold, London, 1991; (*First* edition: *Elements of Sampling Theory.* The English Universities Press Ltd., London, 1974).

2. Cochran, W. G. *Sampling Techniques.* John Wiley & Sons, Inc., New York, 3rd Edition, 1977.

8. Appendix 1: Population Elements and Population Units

As discussed in Section 1 on page 8.3 in Figure 8.1, we distinguish:

- **Elements:** the entities that make up a population; for example, a person is an element of the population of Canadians, but we recognize that many populations in data-based investigating have non-human or *in*animate elements.
- **Units:** the entities *selected* for the sample; a unit may be one element (*e.g.*, a person) or *more than* one (*e.g.*, a household).

This Figure 8.11 is concerned with (survey) *sampling* and so refers in most places to units, but population attributes of interest (like $\mathbf{N}$, $\overline{\mathbf{Y}}$, ${}_{\mathrm{T}}\mathbf{Y}$ and $\mathbf{S}$) refer to *elements.* In introductory courses like STAT 220 and STAT 231, we restrict attention to units which are elements so the distinction is of no consequence but, anticipating Figures 2.14 and 2.16 in STAT 332, when units are *groups* of elements (as in *cluster* sampling), some expressions in the theory must be modified. This is illustrated in Table 8.11.8 at the upper right of the facing page 8.57 by comparing expressions in this Figure 8.11 with those for selecting *equal*-sized clusters, like cardboard boxes in a supermarket that each contain, say, 24 cans of soup, or cartons from a component manufacturing

Figure 8.11. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating an Average or a Total (continued 4)

process that each contain a set number (say, 10) of the component. Table 2.3.7 at the right gives additional notation we need, where the repondent population is considered as $\mathbf{N}$ elements of which n are selected (by EPS) for the sample, or as $\mathbf{M}$ clusters each of L elements, of which m are selected (by EPS) to yield a sample of mL elements.

Given the respondent population 'models' of $\mathbf{N}$ elements or $\mathbf{M}$ clusters, the structural similarity of corresponding expressions in the two columns of Table 8.11.8 are clear; noteworthy points are:

- when estimating $\overline{\mathbf{Y}}$, $\overline{y}$ involves *element* responses y_j but $\overline{y}_{ec}$ involves cluster *average* responses $\overline{y}_j$;
- $\mathbf{S}_{ec}$ (estimated by s_{ec}), which quantifies variation of *cluster averages* in the respondent population, is to be distinguished from the variation of *element* responses quantified by $\mathbf{S}$ (estimated by s).

The cluster sampling expressions in the right-hand column of Table 8.11.8 are taken from Figure 2.14 of the STAT 332 Course Materials. The theory for *un*equal-sized clusters is more complicated – see Figure 2.16 of the STAT 332 Materials.

Table 8.11.7:

	Elements	Clusters	Relationships
Respondent population	$\mathbf{N}$	$\mathbf{M}$	$\mathbf{N} = \mathbf{M}\text{L}$, $\text{L} = \mathbf{N}/\mathbf{M}$
Sample	n	m	$\text{n} = \text{mL}$, $\text{L} = \text{n}/\text{m}$

Also: $\overline{\mathbf{Y}}_i = \frac{1}{\text{L}}\sum_{k=1}^{\text{L}}\mathbf{Y}_{ik}$ is the average response of the ith population cluster,
$\overline{y}_j = \frac{1}{\text{L}}\sum_{k=1}^{\text{L}} y_{jk}$ is the average response of the jth sampled cluster,
the subscript *ec* in Table 8.11.8 below denotes 'equal-sized clusters'.

Table 8.11.8

EPS of elements	Page	EPS of clusters
$\overline{y} = \frac{1}{\text{n}}\sum_{j=1}^{\text{n}} y_j$	8.52	$\overline{y}_{ec} = \frac{1}{\text{m}}\sum_{j=1}^{\text{m}} \overline{y}_j$
$s = \sqrt{\frac{1}{\text{n}-1}\sum_{j=1}^{\text{n}}(y_j - \overline{y})^2}$	8.52	$s_{ec} = \sqrt{\frac{1}{\text{m}-1}\sum_{j=1}^{\text{m}}(\overline{y}_j - \overline{y}_{ec})^2}$
$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^2}$	8.52	$\mathbf{S}_{ec} = \sqrt{\frac{1}{\mathbf{M}-1}\sum_{i=1}^{\mathbf{M}}(\overline{\mathbf{Y}}_i - \overline{\mathbf{Y}})^2}$
$s.d.(\overline{Y}) = \mathbf{S}\sqrt{\frac{1}{\text{n}} - \frac{1}{\mathbf{N}}}$	8.53	$s.d.(\overline{Y}_{ec}) = \mathbf{S}_{ec}\sqrt{\frac{1}{\text{m}} - \frac{1}{\mathbf{M}}}$
$\overline{y} \pm {}_{\alpha}t^*_{\text{n}-1} s\sqrt{\frac{1}{\text{n}} - \frac{1}{\mathbf{N}}}$	8.54	$\overline{y}_{ec} \pm {}_{\alpha}t^*_{\text{m}-1} s_{ec}\sqrt{\frac{1}{\text{m}} - \frac{1}{\mathbf{M}}}$

9. Appendix 2: Representative Sampling

The appealing intuitive idea of a 'representative sample' – one that 'looks like' the (respondent) population with respect to the attribute(s) of interest – is equivocal statistically for four reasons:

- a sample selected by EPS is unlikely to be 'representative' in the sense just given for *all* attributes of potential interest – for instance, a sample may have small [possibly (close to) zero] sample error for estimating $\overline{\mathbf{Y}}$ but large sample error for estimating $\mathbf{S}$;
- the sample, of itself, provides no information about its 'representativeness';
- there is no selecting process known to yield a 'representative' sample, except taking a census;
- the terminology tends to obscure the distinction between the individual case (the *particular* sample) and behaviour under repetition (the properties of the selecting *process*).

Less equivocal terminology is *representative sampling*, with its implication of a selecting *process* (like EPS) which, in conjunction with adequate replicating, provides for quantifying sampling imprecision and so allows a particular investigation to obtain an Answer with acceptable limitation (in the Question context) due to sample error. However, the writer's preference is to *avoid* in statistics the terms 'representative' and 'representativeness' in relation to a sample (or a sampling protocol).

- Kruskal and Mosteller devote 50 pages to discussing the (sometimes ill-defined) meanings in statistical contexts of **representative sampling** in three articles in the *International Statistical Review*, **47**, 13-24, 111-127, 245-265 (1979). [UW Library call number HA 11.I505]

NOTE: 20. An illustration, involving bivariate data, of another instance of sample-attribute dependence is:

- when estimating the least squares *slope* of a straight-line relationship, sample points more concentrated near the *ends* of the interval of observation will reduce sampling imprecision (although this will *in*crease imprecision of any inference needed to show that the relationship *is* a straight line);
- similar considerations apply when estimating *correlation*, although estimating this attribute is rarely discussed.

10. Appendix 3: The Mean of S, $E(S)$

The justification in equation (8.11.17), at the bottom of page 8.53, for using s to estimate $\mathbf{S}$ is compromised by the fact that S is *not* an unbiased estimator of $\mathbf{S}$. Because of the square root in the expression for s in equation (8.11.3) on page 8.49 (and in Table 8.11.5 on page 8.52), there is no simple expression for the estimating bias of the corresponding random variable S under EPS, but we know that bias exists from the following argument, which is an illustration of Jensen's Inequality and uses the fact that the variance of any (non-constant) random variable is positive. We have:

$$0 < var(S) = E(S^2) - [E(S)]^2 = \mathbf{S}^2 - [E(S)]^2 \quad \text{so that, taking square roots:} \quad E(S) - \mathbf{S} < 0. \qquad \text{-----(8.11.24)}$$

NOTE: 21. For the model (8.11.2) on page 8.49, the mathematics is more tractable and leads to equation (8.11.25) at the right so that, because of equation (6.3.38) on page 6.29 of Figure 6.3, rewritten at the right as equation (8.11.26), the bias term multiplying σ on the RHS of equation (8.11.25) is the *mean* of a $K_{\text{n}-1}$ distribution. Table 6.3.10 of its values for n = 2 to 51 (*i.e.*, for 1 to 50 degrees of freedom) on page 6.31 of Figure 6.3 reminds us that estimating bias:

$$E(\tilde{\sigma}) = \sqrt{\frac{2}{\text{n}-1}}\,\frac{\Gamma(\frac{\text{n}}{2})}{\Gamma(\frac{\text{n}-1}{2})}\,\sigma \qquad \text{-----(8.11.25)}$$

$$\frac{\tilde{\sigma}}{\sigma} \sim K_{\text{n}-1} \quad \text{or:} \quad \tilde{\sigma} \sim \sigma K_{\text{n}-1} \qquad \text{-----(8.11.26)}$$

(*continued overleaf*)

NOTE: 21. • decreases in magnitude with increasing sample size, *un*like (real-world) inaccuracy;
(cont.) • of S as an estimator of $\mathbf{S}$ is likely to be unimportant practically for the sample sizes used in most real sample surveys.

11. Appendix 4: Bias and Rms Error

For a random variable Y and some constant c, we have:

$$E\{[Y-c]^2\} = E\{[E(Y)-c+Y-E(Y)]^2\} = E\{[E(Y)-c]^2+[Y-E(Y)]^2+2[E(Y)-c][Y-E(Y)]\}$$
$$= E\{[E(Y)-c]^2\}+E\{[Y-E(Y)]^2\}+2E\{[E(Y)-c][Y-E(Y)]\}$$
$$= [E(Y-c)]^2+E\{[Y-E(Y)]^2\}+2[E(Y)-c]E[Y-E(Y)]$$

i.e., $E\{[Y-c]^2\} = [E(Y-c)]^2+[s.d.(Y)]^2$ because $E[Y-E(Y)] \equiv 0.$ -----(8.11.27)

If we now think of Y as a random variable whose distribution represents the possible values of a *response variate* $\mathbf{Y}$ and c as a *true* value, the left-hand side of equation (8.11.27) is a *mean squared error* and $E(Y-c)$ in the first term on the right-hand side is a *bias*; we can therefore interpret equation (8.11.27) as:

mean squared error = bias² + standard deviation². -----(8.11.28)

Taking the square root so we are working on the *same* scale as the variate represented by Y, the **root mean squared error** is:

rms error = $\sqrt{\text{bias}^2 + \text{standard deviation}^2}$. -----(8.11.29)

Thus, the rms error is *one* concept that *combines* the two model quantities of bias and (probabilistic) standard deviation, corresponding to the two real-world entities of inaccuracy and imprecision.

Equation (8.11.29) provides useful insights about bias and variation in the context of survey sampling; different cases depend on how broad our focus is in terms of *which* true value c represents – see also the discussion and diagram showing four components of overall *error* on the lower half of page 5.25 in Figure 5.7 of the STAT 231 Course Materials.

* The narrowest focus is *measuring* when c is the true value of the response variate $\mathbf{Y}$; equation (8.11.29) is then:

  measuring rms error = $\sqrt{\text{measuring bias}^2 + \text{measuring standard deviation}^2}$. -----(8.11.30)

* For measuring *and* sampling, c is the true value of the *respondent* population attribute of $\mathbf{Y}$ and then:

  measuring *and* sampling rms error = $\sqrt{\text{measuring + sampling bias}^2 + \text{measuring and sampling standard deviation}^2}$; -----(8.11.31)

NOTE: 22. Measuring and sampling standard deviation = $\sqrt{\text{measuring standard deviation}^2 + \text{sampling standard deviation}^2}$. -----(8.11.32)

* For measuring *and* sampling *and* non-responding, c is the true value of the *study* population attribute of $\mathbf{Y}$ and then, under our assumption that non-response is *deterministic* (*not* stochastic):

  measuring and sampling and non-responding rms error = $\sqrt{\begin{array}{l}\text{measuring + sampling}\\ \text{+ non-responding bias}^2\end{array} + \text{measuring and sampling standard deviation}^2}$. -----(8.11.33)

* For measuring *and* sampling *and* non-responding *and* specifying, c is the true value of the *target* population attribute of $\mathbf{Y}$ and then, under our assumption that specifying the study population also is *deterministic*:

  measuring and sampling and non-responding and studying rms error = $\sqrt{\begin{array}{l}\text{measuring + sampling}\\ \text{+ non-responding}\\ \text{+ studying bias}^2\end{array} + \text{measuring and sampling standard deviation}^2}$. -----(8.11.34)

NOTE: 23. In printed materials other than these Course Materials (*e.g.*, see Cochran, p. 15), equation (8.11.27) [or (8.11.28)] is usually discussed only with respect to *estimating* bias. Although estimating bias is a relatively minor topic in STAT 220, it is useful to recognize the following [recall also Example 8.11.1 on pages 8.51 and 8.52]:

- **Estimating bias** (a *model* quantity) is the difference between the mean of an estimator and the value of the corresponding population attribute (or model parameter); for example, under EPS:
  - the random variable $\overline{Y}$ representing the sample *average* $\overline{y}$ is an *un*biased estimator of the respondent population average $\overline{\mathbf{Y}}$ because, as shown in equation (8.11.8) at the top of the fifth side (page 8.53) of this Figure 8.11, $E(\overline{Y}) = \overline{\mathbf{Y}}$ or $E(\overline{Y}) - \overline{\mathbf{Y}} = 0$; BUT
  - the sample ratio $r = \overline{y}/\overline{x}$ is a *biased* estimator of the respondent population ratio $\mathbf{R} = \overline{\mathbf{Y}}/\overline{\mathbf{X}}$ because $E(R) \neq \mathbf{R}$ or $E(R) - \mathbf{R} \neq 0$, and likewise for S as an estimator of $\mathbf{S}$ as discussed overleaf on page 8.57 in Appendix 3.
- The rms error of an estimator is of interest because, while we prefer an *un*biased estimator of a population attribute, there are times when a *biased* estimator has only *small* bias and appreciably *smaller* standard deviation than an available *un*biased estimator; we *may* then prefer the biased estimator with *smaller* rms error.
- *Un*like (real-world) inaccuracy, estimating bias *de*creases in magnitude with increasing sample size (as discussed in Appendix 3 overleaf on page 8.57 and above in Note 21).